Workshop: Natural language and AI

Nicholas Catasso (Bergische Universität Wuppertal), Thomas Scharinger (Friedrich-Schiller-Universität Jena)

New perspectives for linguistic studies In recent years, Artificial Intelligence (AI) has made significant strides across various disciplines. The integration of AI applications, particularly Large Language Models (LLM), into linguistic studies has opened new horizons for the analysis of natural language. From morphosyntax to semantics and variation linguistics, these technologies provide linguists with the opportunity to explore complex linguistic phenomena. This development has led to the automation of linguistic tasks such as text generation, translation, and corpus annotation to an extent that was previously unimaginable. However, the application of AI in linguistic research also reveals challenges. One significant issue lies in the need to provide adequate training data for AI models, covering a wide range of linguistic phenomena and structures. Often, these data are incomplete, uneven, or even erroneous, which can compromise the reliability of AI systems. Another obstacle is the fact that AI models may inherit implicit biases from the existing data on which they are trained. This can result in distortions in the results and compromise the neutrality and objectivity of linguistic analyses. Furthermore, language variation seems to pose a challenge. AI models must be able to recognize and process this diversity appropriately. This is often difficult as the models may be constrained by certain linguistic patterns or norms. The planned workshop will be a forum to discuss how and to what extent AI applications (such as ChatGPT, DialoGPT, Meena, BlenderBot, etc.) may be relevant for linguistic studies. By merging theoretical approaches in linguistics with modern AI methods, the potentials and challenges of these technologies for linguistic research will be explored. The workshop will focus on – but will not be limited to – the following research questions:

• How can AI applications be used to investigate and compare grammaticality in different languages? Which applications come closest to native speaker intuition?
• To what extent can Large Language Models (LLM) be used for automatic corpus annotation and analysis to identify and understand linguistic variation?
• What role do semantic models and neural networks play in translating between languages with different syntactic structures?
• How can Natural Language Processing (NLP) techniques be used to examine the meaning and usage of language variation in different social contexts?

Alessia Battista
(Università degli Studi di Napoli "Parthenope" & Università degli Studi di Salerno)

Can AI assist human researchers? A sample genre analysis using ChatGPT

Considering the evolving nature of business communication, particularly as influenced by the recent digital media and the impact of technology in general, this study will try to understand whether *BuzzFeed*'s *Tasty*' has been contributing to the rise of new discursive genres in online business communication. This contribution investigates the influence of social media networks on the rhetorical structure of *Tasty*'s recipe videos posted on Facebook, Instagram, and TikTok in May 2023. The analysis relies on Cesiri's (2020) framework for the genre analysis of food blogs, which has been expanded for the study of social media platforms. Additionally, the analysis is performed both manually by the human researcher and through a custom GPT model named *Genre Analyst*, which has been developed to identify rhetorical moves and variations. Comparing the human and the Ai analyses allows to accurately identify genre moves, while also providing additional insights on tone and engagement. The study thus contributes to the digital media research and underscores the potentialities of AI tools in linguistic analysis; nevertheless, research in this sense is still at an early stage, which implies the need to explore larger datasets and further examine if and how AI can support human researchers.

References

Cesiri, Daniela. 2020. *The Discourse of Food Blogs: Multidisciplinary Perspectives*. 1st ed. Routledge. https://doi.org/10.4324/9780429455865.

Nicholas Catasso
(Bergische Universität Wuppertal)

Benchmarking AI grammar: A study of ChatGPT's and human grammaticality judgments

The rapid development of AI technologies, particularly large language models (LLMs) like ChatGPT-4, has transformed linguistic research, enabling advanced processing and analysis beyond traditional methods (cf. i.a. Austin et al. 2021; Liu et al. 2023; Torrent et al. 2023; Yu et al. 2024; Qiu et al. 2024). One of the most promising applications of LLMs lies in their ability to generate grammaticality judgments, a core concept in theoretical syntax. Grammaticality refers to whether a sentence conforms to the structural rules of a language, traditionally evaluated through the intuitive judgments of native speakers. Grammaticality itself, however, is not always binary; it has been shown to exhibit gradient acceptability, depending on various linguistic and contextual factors (Schütze 2016; Sprouse 2018). While LLMs offer new opportunities to model these judgments, the degree to which they can replicate or approximate the nuanced and sometimes variable human assessments remains underexplored. This pilot study explores the capacity of ChatGPT 4 to produce grammaticality judgments for German sentences, comparing its performance to that of native German speakers. The research addresses two main questions: (i) How well does ChatGPT-4 replicate the grammaticality judgments typically made by humans?; (ii) what factors contribute to any observed differences between the judgments generated by the AI and those provided by human participants? To investigate these issues, both ChatGPT-4 and native speakers were presented with the same linguistic stimuli, designed to test a variety of syntactic structures in German. The resulting judgments were then analyzed and compared, offering an assessment of how closely the AI model aligns with human intuitions. The findings of this study provide insights into the strengths and limitations of using LLMs like ChatGPT-4 for linguistic tasks that rely on nuanced syntactic knowledge. While ChatGPT-4 demonstrates considerable accuracy in replicating human judgments, notable discrepancies highlight areas where AI models still fall short of human linguistic competence. These results contribute to the discussion on AI's role in computational linguistics, particularly the reliability of LLMs in handling human-like syntactic reasoning, and pave the way for further advancements in AI-driven linguistic research.

References

Austin, J. et al. 2021. Program synthesis with Large Language Models. arXiv abs/2108.07732.
Liu, H. et al. 2023. Evaluating the logical reasoning ability of ChatGPT and GPT-4. arXiv preprint arXiv:2304.03439
Qiu, Z. et al. 2024. Grammaticality representation in ChatGPT as compared to linguists and laypeople. arXiv preprint arXiv:2406.11116.
Schütze, C. T. 2016. *The empirical base of linguistics. Grammaticality judgments and linguistic methodology*. Berlin: Language Science Press.
Sprouse, J. 2018. Acceptability judgments and grammaticality, prospects and challenges. In N. Hornstein et al. (eds.), *Syntactic Structures after 60 Years. The impact of the Chomskyan revolution in linguistics.* Berlin: de Gruyter, 195-223.
Torrent, T. et al. 2023. *Copilots for linguists: AI, constructions and frames*. Cambridge: CUP.
Yu, D. et al. 2024. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies*. International Journal of Corpus Linguistics* 29/3: 1-28.

Denise Löfflad, Detmar Meurers, Elina Schnaper & Benedikt Beuttler
(Universität Tübingen, Leibniz Institut für Wissensmedien (IWM))

Automatic CEFR-based linguistic analysis for LLM generated texts

Reading is fundamental to second language acquisition [1]. However, finding authentic, engaging, and level-appropriate texts remains a significant challenge, particularly for low proficiency learners. Suitable texts must be grammatically and lexically simple enough to be comprehensible, requiring extremely basic language at the beginner level, which authentic texts rarely provide. As a result, learners often rely on textbook texts that may not cover interesting or current topics.
The release of tools like ChatGPT by OpenAI and other Large Language Models (LLMs) has created new possibilities, making it easy to generate texts tailored to learners' interests across various topics. These models can be used for metalinguistic analysis [2] or as language tutors [3] but are still prone to errors as they can produce false output [4]. Although these models are trained on extensive (but largely unknown) datasets, they lack explicit grounding in linguistic principles and educational theories.
The Common European Framework of Reference (CEFR,[5]) is widely used for proficiency assessment and describes proficiency levels while aiming for adaptability to various European languages. This neutrality ensures comparability of proficiency levels across languages, but simultaneously results in the CEFR omitting specific linguistic constructs crucial for automatically analyzing texts.
Within the POLKE-GER project, we aim to address this gap by developing a tool designed to automatically extract pedagogically significant linguistic knowledge, drawing on CEFR and Profile Deutsch descriptors [6]. We use this tool to analyze German texts, identifying linguistic structures relevant to the CEFR. Additionally, we employ prompt engineering techniques to generate texts using LLMs, which we then analyze with our tool to assess their alignment with the appropriate linguistic constructs.
While still in development, our tool shows promising results with 33 currently implemented features, demonstrating high precision and recall (mean *p=0.92*, mean *r=0.82*). Initial experiments with ChatGPT reveal the possibility to simplify existing structures through prompting, but an overall inability of LLMs to produce texts fully aligned with learners' proficiency levels, as ChatGPT often introduces new linguistic constructs inappropriate for the target proficiency level. Future work will focus on more exhaustive feature exploration and integrating our tool into existing platforms for easy usability.
Our approach allows users to quickly identify potentially difficult linguistic constructs. This enables both learners and teachers to use LLM-generated texts more productively, reducing frustration and helping to tailor materials more effectively to individual learners. Ultimately, our goal is to create a hybrid model that leverages both traditional computational linguistics and modern LLM techniques to enhance educational AI systems.

References

[1] S. D. Krashen, *The Power of Reading: Insights from the Research*. Bloomsbury Publishing USA, 2004. [1]
[2] G. Beguša, M. Dąbkowskia, and R. Rhodesb, 'Large linguistic models: Analyzing theoretical linguistic abilities of LLMs', *arXiv preprint*, 2023.
[3] F. Karataş, F. Y. Abedi, F. O. Gunyel, D. Karadeniz, and Y. Kuzgun, 'Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners', *Springer*, pp. 1–24, 2024.
[4] J. Kocoń *et al.*, 'ChatGPT: Jack of all trades, master of none', *Inf. Fusion*, vol. 99, p. 101861, Nov. 2023, doi: 10.1016/j.inffus.2023.101861.
[5] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.
[6] M. Glaboniat, P. Rusch, H. Schmitz, and L. Wertenschlag, *Profile Deutsch*, vol. 21. Berlin: Langenscheidt, 2002.

Pavel Grashchenkov, Lada Pasko & Kseniia Studenikina
(Lomonosov Moscow State University)

RuParam: Russian corpus of linguistic minimal pairs (ONLINE)

The problem of assessing the linguistic competence of language models (LMs) has been widely discussed in recent years. The state-of-the-art sources for evaluation, CoLA (Warstadt et al. 2019) and BLiMP (Warstadt et al. 2020), are based on the linguistic notion of grammaticality. These benchmarks are aiming to test if the LMs are able to distinguish between the sentences that follow the grammar rules and the ones that fail to do so. Data in CoLA originates from theoretical linguistic literature; each sentence is labeled as either grammatical or ungrammatical. In contrast, BLiMP is a corpus of minimal pairs where each grammatical sentence is paired with one containing a mistake; the data were generated artificially. Both of these corpora were developed for English. Given that LMs are available for a wide range of languages, tools for linguistic competence evaluation should account for this diversity. Since grammar rules are to a large extent language-specific, corpora cannot be generated by mere translation and should be designed for any given language specifically. In our talk, we will be focusing on the grammaticality corpus we have created for Russian.

A solution to the problem of linguistic evaluation for Russian has already been proposed by RuCoLA (Mikhailov et al. 2022), a corpus created in accordance with the CoLA methodology. However, RuCoLA seems to have certain weakpoints in terms of linguistic adequacy. Firstly, many phenomena covered by RuCoLa show variation, thus the (un)grammaticality contrast becomes irrelevant for them. Secondly, RuCoLA provides only a very broad classification of ungrammaticality sources (e.g. 'syntax', 'semantics'), which makes the possible linguistic analysis of LMs competence rather superficial. Finally, the corpus for the most part focuses on complex linguistic phenomena (which are more often discussed in theoretical linguistics literature), underrepresenting basic grammar notions of Russian, such as predicate agreement.

We have developed RuParam, a new linguistic competence evaluation corpus for Russian, aimed at overcoming the discussed issues. Similarly to BLiMP, our database consists of pairs of sentences differing in grammaticality. So far, RuParam contains approximately 8.8k of minimal pairs and consists of two parts. The first part employs a novel source of grammaticality judgement data: we use materials from the Test of Russian as Foreign Language (TORFL). Minimal pairs are automatically retrieved from multiple-choice tasks assessing grammar and vocabulary skills. For each pair, the source of ungrammaticality is manually annotated by linguists using one or a few of 26 labels. As TORFL is designed for different levels of language proficiency (CERF A1–C2), our data include phenomena ranging from basic to complex from the perspective of L2 learners. We suppose that this range may be relevant to the study of multilingual LMs, which have not been much trained on Russian data. Since the original tasks had only one correct answer, our dataset is free of linguistic variation. The second part represents phenomena that are more sophisticated and therefore absent from the TORFL tasks. These include 28 categories, such as island constraints, non-projectivity, anaphor binding, and licensing of negative polarity items. The ungrammatical sentences are manually derived from their counterparts found in real texts.

We tested seven LMs using a prompt requiring the model to select the correct sentence from the pair. We consider this methodology to be the most ecologically valid because, on the one hand, it requires no training and tests the model's linguistic faculty 'as it is'; on the other hand, it examines not only the model's ability to assign a higher probability to grammatical sentences, but also its understanding of grammaticality and correctness. The models we tested show different accuracy rates on our dataset, ranging from 0.60 to 0.93. In our talk, we will take a closer look at both the design of the dataset and the results of the models.

References

Mikhailov, V., Shamardina, T., Ryabinin, M., Pestova, A., Smurov, I., & Artemova, E. (2022). RuCoLA: Russian Corpus of Linguistic Acceptability. *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5207–5227.

Warstadt, A., Singh, A., & Bowman, S.R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, vol. 7, 625–641.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, Sh., & Bowman, S.R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, vol. 8, 377–392.

Preeti Kumaari, Matthew Galbraith & Martina Wiltschko
(Universitat Pompeu Fabra)

Grammar plays a role in Human-Computer Interaction, huh? (ONLINE)

Introduction. Natural language sentences may have two components- propositional (p-language) and interactional (i-language) (Wiltschko 2021). For example, in "The alarm is set for 7 AM, huh?", the propositional content is "The alarm is set for 7 AM", while the interactional "huh?" has two functions: i) it indicates that the speaker assumes some shared knowledge with the addressee, thus regulating the management of common ground (CG); ii) it signals that the speaker requests a response, thus managing turn-taking (TT). Our question is whether human users perceive the use of i-language in Human-computer Interaction (HCI) as natural. Particularly, whether there is a difference between i-language used for CG management and TT management. Our findings suggest that i-language required for CG management is considered unnatural in HCI, whereas its use for TT management gave us inconclusive results. Methodology. An acceptability judgement experiment was conducted through a survey wherein 200 native English speakers assessed the naturalness of HCI employing "huh?" in two contexts: 1) Other Initiated Repair (OIR), a case of TT-management and 2) Request for Confirmation (RFC), a case of CG management. Participants were given five dialogues containing OIR and RFC each. For OIR, participants evaluated two hypothetical HCI scenarios - one, an interrogative formulated with p-language only (OIRp) and the other using i-language "huh?" (OIRi), see Table 1 (where H = human user and C = computer).

Table 1. OIR Dialogue

	OIRp	OIRi
H:	Set an alarm for 8 o'clock at night.	Set an alarm for 8 o'clock at night.
C:	OK, I can do that for you.	OK, I can do that for you.
H:	**What was that?**	**Huh?**

In the RFC scenario, subjects rated the naturalness of four target utterances - a standard interrogative (Interrogative in Table 2), a confirmational question with "huh?" (RFCi in Table 2), a declarative (Declarative in Table 2), and a confirmational question with "Is that true?" (RFCp in Table 2). We used these contexts to see how human users perceive the difference between a canonical interrogative and an RFC question (columns 2 & 3) as well as whether there is a difference in perception between RFCs with and without i-language (columns 2 & 4). Finally, we also tested the naturalness of an RFC using a declarative, a possibility in Human-human Interaction (HHI).

Table 2. RFC Dialogue

	Interrogative	RFCi (i-language)	Declarative	RFCp (p-language)
H to C	Do I have an alarm set for today?	I have an alarm set for today, huh?	I believe that I have an alarm set for today.	I believe that I have an alarm set for today, is that true?

Results. To analyse the collected data, ordinal regression analyses were conducted on the Likert scale responses for each context. For OIRs, we observed a tendency towards p-language, but we did not find a statistically significant result for "huh?" being classified as natural or unnatural. For RFCs, interrogatives were rated significantly natural, whereas RFCi were rated significantly unnatural, with the overall order of naturalness being as follows: Interrogative > RFCp > Declarative > RFCi

Discussion. In OIR dialogues, responses lacking "huh?" showed higher naturalness, but the inconsistent statistical significance suggests further investigation is needed to reach a definitive conclusion. As for RFCs, the result suggests that i-language which regulates CG is considered unnatural in HCI. Thus, one of the core attributes of natural language used in HHI is affected in HCI. This is not surprising as i-language is highly context-specific. We submit that i-language can be used as a window into the way

human users view computers as interactants without having to rely on post-hoc questionnaires (Bartneck et al 2009).

References

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of Robots. International Journal of Social Robotics, 1(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Clark HH, & Fischer K. (2022). Social robots as depictions of social agents. Behav Brain Sci. Mar 28;46:e21. doi: 10.1017/S0140525X22000668. PMID: 35343422.

Dingemanse, M., Torreira, F., & Enfield, N. J. (2013). Is "Huh?" a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items. *PLoS ONE,* 8(11), e78273. https://doi.org/10.1371/journal.pone.0078273

Wiltschko, M. (2021). *The Grammar of Interactional Language* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108693707

Wiltschko, M., & Heim, J. (2016). The Syntax of Confirmationals. A neo-performative analysis. In G. Kaltenböck, E. Keizer, & A. Lohmann (Eds.), *Outside the clause: Form and function of extra-clausal constituents* (pp. 303–340). John Benjamins.

Gohar Rahman
(Islamia College, Peshawar)

Automating semantic annotation in understudied languages: A case study of Urdu corpus using GPT-4 (ONLINE)

This study explores the application of a fine-tuned GPT-4 model for automating the annotation of a semantic corpus in Urdu, an underrepresented language in Natural Language Processing (NLP).
Traditionally, corpus annotation has been a labor-intensive and time-consuming task, especially for low-resource languages like Urdu, which lack extensive pre-existing linguistic datasets. This research addresses the gap by leveraging the capabilities of Large Language Models (LLMs) to automate the process, focusing on annotating semantic roles such as agents (فاعل), actions (عمل), and objects (مفعول) within Urdu sentences. We created a manually annotated Urdu corpus, which was then used to fine-tune GPT-4, evaluating the model's performance using precision, recall, and F1-score metrics.
The model's overall performance showed considerable promise, with precision at 85%, recall at 83%, and an F1-score of 84%. These results suggest that AI-driven annotation approaches can achieve near-human accuracy when fine-tuned on domain-specific linguistic data. Moreover, the automation significantly reduced the time required for annotation, offering a scalable solution for the development of linguistic resources in low-resource languages. In particular, the model excelled at identifying verbs (actions) and nouns (agents and objects), demonstrating a robust ability to handle core semantic roles. However, challenges arose in tagging prepositions (e.g., کو, سے) and conjunctions (e.g., اور, لیکن), which require further refinement or the inclusion of supplementary rule-based systems. This discrepancy indicates the complexity of Urdu's syntactic structure, where postpositions and conjunctions often bear significant semantic weight, complicating the model's interpretation.
The findings of this research highlight the transformative potential of AI in linguistic annotation, especially for underrepresented languages like Urdu, where annotated corpora are limited. The successful application of the GPT-4 model in this context suggests broader implications for NLP research, particularly in enhancing the accessibility of computational resources for low-resource languages. By automating annotation processes, AI not only reduces manual labor but also improves consistency and scalability, making it a valuable tool for linguistic studies in languages lacking substantial NLP infrastructure. Furthermore, this study emphasizes the importance of fine-tuning LLMs on language-specific data to optimize accuracy and adaptability. As NLP increasingly incorporates diverse linguistic datasets, the inclusion of underrepresented languages like Urdu will contribute to a more inclusive, culturally representative body of research. The success of this approach offers promising opportunities for future research to improve the annotation of more complex linguistic features, such as discourse markers and pragmatic elements, further enhancing AI's capacity to handle the intricacies of human language. Future work could also explore integrating rule-based approaches to refine the model's understanding of challenging syntactic structures, thereby improving accuracy and expanding its application in more complex linguistic tasks.

Matthias Schöffel[1] & Marinus Wiedner[2]
([1]Bayerische Akademie der Wissenschaften, [2]Universität Freiburg)

Simulating the development of Grammatical Gender from Latin to Old Occitan

This communication is based on the study by Polinksy/Everbroeck (2003), who simulated the reanalysis and reattribution of grammatical gender from Latin to Old French with a connectionist model. Building on this, we want to simulate the development of gender from Latin to Old Occitan with a character-based approach. To this end we use a Long-short-Term Memory (LSTM) architecture combined with an attention mechanism in opposition to heuristic models (cf. Marr/Mortensen 2020).
A gender reduction from three to two genders took place in the transition from Latin to Old Occitan, during which the neuter disappeared. The neuter nouns (e.g. Latin third declension MARE) had to be reattributed to either masculine or feminine (cf. it. *il mare*$_{masc}$ vs. fr. *la mer*$_{fem}$ vs. both genders in Old Occitan), and we aim at simulating this development, starting on the character level.
For our current computer simulation, we use nouns from the *Dictionnaire de l'occitan medieval* (DOM), the largest work of Old Occitan lexicology. We also included variants that had to be digitized beforehand via a tailored OCR model (cf. Garcés Arias/Pai/Schöffel/Heumann/Aßenmacher). As a starting point for the model training we take the linked etyma from the *Französischen Etymologischen Wörterbuch* (FEW).
In addition to the lexicographic (and therefore normalised) data we use nouns extracted from original manuscripts from the 13th and 14th century, semi-automatically transcribed with a Transkribus-model for Old Occitan Handwriting (cf. Wiedner 2023). We then annotated the texts by using available PoS tagger, and manually correcting the results (which, in addition, allows us to compare these taggers regarding their accuracy). Afterwards, we manually combined these nouns with their respective etyma, including information on gender (including possible variation) and the accusative forms; the required information is taken from the FEW and the *Thesaurus Linguae Latinae* (TLL). We want to see if there are differences in the simulations' outcome with this non-normalised, 'authentic' data in comparison to the data taken from the DOM.
We will present and discuss the basic idea as well as preliminary results.

References

DOM = Dictionnaire de l'occitan médiéval. <http://www.dom-en-ligne.de/>.
FEW = Wartburg, Walther von, et al. (1922–2022): Französisches Etymologisches Wörterbuch (FEW). Eine Darstellung des galloromanischen Sprachschatzes. 25 Bände, Bonn/Heidelberg/ Leipzig/ Berlin/ Basel, Klopp/ Winter/ Teubner/ Zbinden. <https://apps.atilf.fr/lecteurFEW/>
Garcés Arias, Esteban, Pai, Vallari, Schöffel, Matthias, Heumann, Christian, Aßenmacher, Matthias. 2023. Automatic Transcription of Handwritten Old Occitan Language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics, 15416–15439.
Marr, Clayton, Mortensen David (2020): „Computerized Forward Reconstruction for Analysis in Diachronic Phonology, and Latin to French Reflex Prediction", *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, 28–36.
Polinsky, Maria, Everbroeck, Ezra van (2003): „Development of Gender Classifications: Modeling the Historical Change from Latin to French", *Language* 79, 356–390.

TLL = Thesavrvs linvae Latinae (TLL) open access. München. <https://publikationen.badw.de/de/thesaurus>
Wiedner, Marinus (2023). OldOccitanHandwriting, (Modell-Nr. 52822, CER=3,51 %), PyLaia-model for handwritten Occitan of the 13th and 14th century. <https://readcoop.eu/de/modelle/old-occitan-handwriting/>

Petra Sleeman
(Universiteit van Amsterdam)

AI translations of Germanic bare plural subjects into Romance

In Germanic, bare plural subjects can be used with an existential and a generic interpretation (Carlson 1977; Longobardi 1994). Delfitto & Schroten (1991) judge existential bare plurals in subject position acceptable in English and Dutch, but not in Spanish and Italian. Giusti (2021) also makes a distinction between Germanic and Romance. According to Dobrovie-Sorin & Laca (2003) as well, existential bare plural subjects are excluded in (European) Romance languages. French does not have bare nouns, but has indefinite plural nouns introduced by a partitive article: *des*. Dobrovie-Sorin & Laca (2003) claim that, in French, existential subjects introduced by the partitive article have the same semantic and pragmatic properties as the English ones, which would make them acceptable in preverbal subject position (see also Bosveld-de Smet 2004 and Ihsane 2008). For languages in which existential bare plural subjects have been claimed to be unacceptable, it has been observed that an additional adjective or PP or narrow or contrastive focus may make the bare subject acceptable (Longobardi 1994 for Italian; Suñer 1982, Salem 2010, Leonetti 2013 for Spanish; Müller & Oliveira 2004 for European Portuguese). In contrast to Germanic plural subjects with a generic interpretation, which are bare, Romance generic plural subjects are introduced by a definite article. Not only plural subjects with an existential interpretation, but also plural subjects with a generic interpretation may be used bare when a modifier is added (Mari 2017). Longobardi (2002) shows, however, that this is not possible with nouns with kind-level predicates.

To research the influence of an existential versus a generic interpretation, modification and the type of predicate on the choice of a determiner, I investigated how AI-powered translators translate Germanic sentences with bare plural subject nouns into French, Italian, Spanish and European Portuguese I used three translators: Google Translate, DeepL and ChatGPT. I selected 108 bare plural subject nouns from two novels by the Dutch author Ilja Leonard Pfeijffer and divided them into four groups: non-modified existential (29 bare plural subjects), modified existential (40), non-modified (21) and modified generic/kind (18). I submitted the Dutch sentences in their context to the translators. In the analysis I distinguished the four types of nouns, the four languages and the three types of digital translators. One of the novels has been translated by human translators into the four Romance languages, the other one only into Italian. I also analyzed these official translations.

The results show that modification largely enhances the use of bare/*des* existential subjects, especially in Italian and Spanish. ChatGPT used many non-modified existential *des*-nouns in French and even more bare nouns in Portuguese, but did not exclude bare nouns in the Italian and Spanish translations. Of the three digital translators, ChatGPT reflected most the translations by the human translators. The three digital translators almost exclusively used the definite article with non-modified generic nouns, but ChatGPT also used bare nouns in the Portuguese translations, which is not excluded according to Brito & Lopes (2016). As for the 18 modified generics/kinds, on the basis of the results I distinguished two subgroups: 10 definite and 8 indefinite generics/kinds. All translators were rather consistent in their use of a definite or an indefinite article for the two subgroups. ChatGPT used again relatively much more bare nouns for Portuguese in the definite subgroup.

The analysis suggests that AI-powered translators may help to confirm, reject or refine claims made in the linguistic literature. Translations of the same texts into different languages may reveal subtle differences between the languages, but also inconsistencies in the translations.

References

A.-M. Brito & R. Lopes (2016). The structure of DPs. In *Handbook of Portuguese Linguistics*.
G. Longobardi (2002). How comparative is semantics? In *Natural Language Semantics 8*.
A. Mari (2017). Pour une anatomie des règles. In *Approches plurielles du nom sans déterminant*.

Franz Meier & Martha Kaiser
(Universität Augsburg)

Turn management in ChatGPT-generated French and Spanish conversations – A case of fictional orality?

The aim of this contribution is to investigate the ability of ChatGPT to generate French and Spanish conversations. The focus is on informal conversation situations of communicative immediacy. The study focuses on features of interaction-in-talk and aims at determining how ChatGPT models turn-taking which, in fact, is constitutive for the progressivity of conversations. In this context, it is not only of interest how speaker change is organized in ChatGPT-generated conversations in comparison to real or 'natural' conversations, but also which differences and similarities arise with human-scripted fictional conversations, such as in films or theatre plays (cf. Kerbrat-Orecchioni 1996; Herman 1998; Bednarek 2010). While in 'natural' conversations the distribution of speaking rights develops out of the conversation itself, the organization of speaker change in fictional conversations is different: "speaker change is not locally managed but totally author-controlled with turn-taking rights being established on dramaturgical grounds rather than on democratic conversational principles" (Spitz 2005, 22). Based on communicative contexts from French and Spanish film scripts, ChatGPT was asked to generate different conversations (e.g. disputes) as a) natural conversations and b) scripted conversations. The investigation of these AI-generated conversations focuses on the structure and distribution of turns from a conversation analytical perspective (Sacks/Schegloff/Jefferson 1974; Auer 2020) and is carried out in contrast to scripted conversations written by humans. With this experimental setup, it can be investigated if ChatGPT-generated conversations show different degrees of fictional orality (cf. Dufter/Hornsby/Pustka 2020).

References

Auer, Peter. 2020. „Die Struktur von Redebeiträgen und die Organisation des Sprecherwechsels." *Einführung in die Konversationsanalyse*. Herausgegeben von Karin Birkner, Peter Bauer, Angelika Baur und Helga Kotthoff. De Gruyter.

Bednarek, Monika. 2010. *The language of fictional television. Drama and identity*. Continuum.

Dufter, Andreas, David Hornsby, und Elissa Pustka. 2020. „L'oralité mise en scène dans la littérature: aspects sémiotiques et linguistiques." *Zeitschrift für französische Sprache und Literatur* 130, 1. 2-19.

Herman, Vimala. 1998. „Turn-management in Drama." Exploring the Language of Drama: From Text to Context. Herausgegeben von Jonathan Culpeper, Mick Short, Peter Verdonk. Routledge.

Kerbrat-Orecchioni, Catherine. 1996. „Dialogue théâtral vs conversations ordinaires." *Cahiers de praxématique* 26.

Sacks, Harvey, Emanuel Schegloff, und Gail Jefferson. 1974. „A Simple Systematic for the Organisation of Turn Taking in Conversation." *Language* 50, 4. 696-735.

Spitz, Alice. 2005. *Power Plays. The representation of mother-daughter disputes in contemporary plays by women. A study in discourse analysis*. Saarland University.